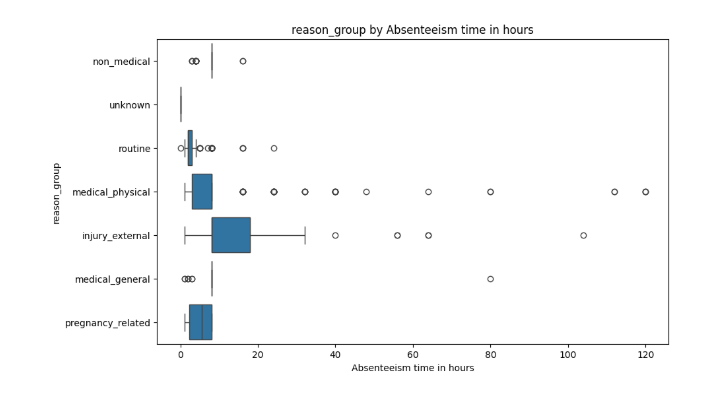
**Interpreting the UCI Absenteeism dataset**

**Overview**

Initially, the data was subdivided into 29 reasons for absence. In order to make sense of a small dataset with many features, the reasons for absence were clustered as shown below. This table includes median and mean hours absent per instance, total frequency, and the standard deviation of the hours absent.

*Table 1: Absenteeism in hours*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Groups | Median | Mean | Frequency | STD |
| Injury external | 8 | 18.2 | 40 | 22.1 |
| Medical general | 8 | 9.3 | 27 | 14.4 |
| Medical physical | 8 | 13.4 | 189 | 21 |
| Unjustified | 8 | 7.5 | 71 | 2.2 |
| Pregnancy related | 5.5 | 5 | 6 | 3.3 |
| Routine checkups | 2 | 2.9 | 364 | 2.4 |
| Disciplinary action | 0 | 0 | 40 | 0 |



*Figure 1: Box plot of the different groups of absence reasons*

**Modelling various factors**

By attempting to predict disciplinary failure (getting fired), it quickly becomes evident that the amount of previous absences is a strong predictor. As seen in figure 2, previous absences is the clear dominant feature for predicting disciplinary failure. Besides that, a high amount of routine checkups and higher travel costs are also correlated to disciplinary failure.

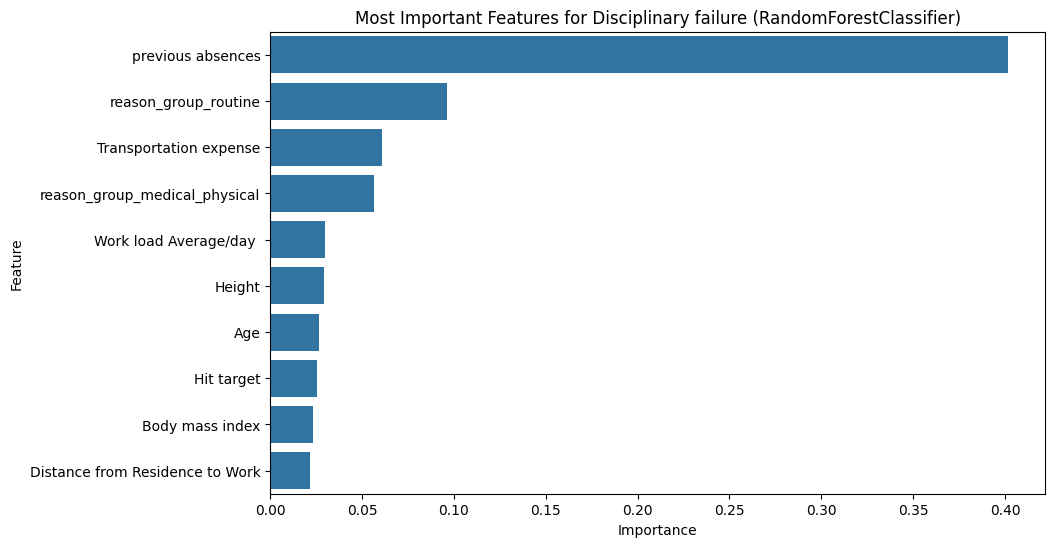
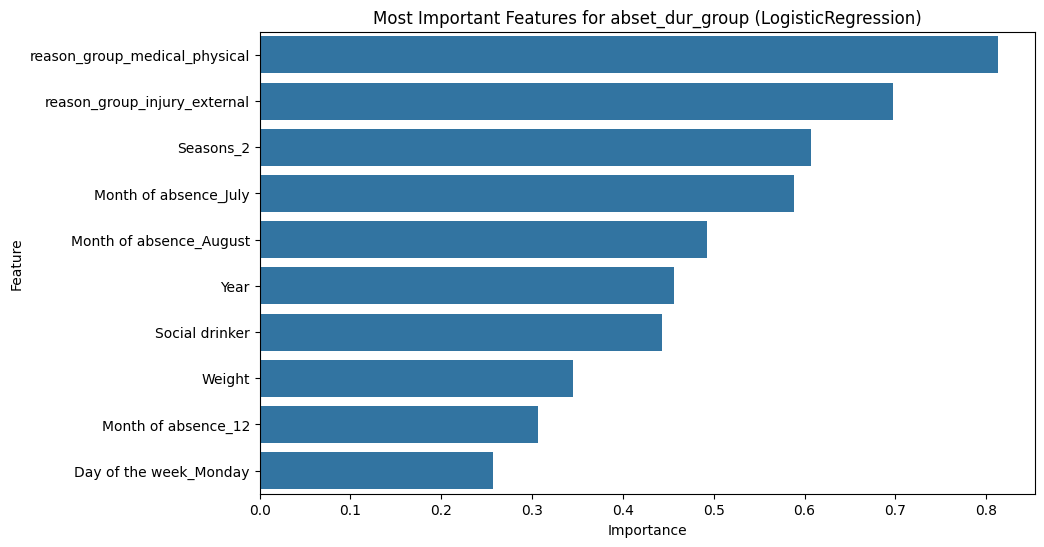
*Figure 2: Disciplinary failure modelled*

Figure 3 shows that the two groups most strongly correlated with longer absences are external injuries and physical medical illnesses. This makes sense, given that these are the most debilitating of the injuries. Absences in season 2 (summer) and the months of July and August are likely to be related to holidays. Lastly, weight and social drinking seem to be related to longer absence, most likely because of their negative correlation with general health. Encouraging healthy dieting and movement practices as well as discouraging (heavy) drinking events related to work could lead to a significant decrease in absenteeism due to medical issues.

*Figure 3: Long absences modelled*



It’s important to note that these are all correlational findings, not causational. Further research is needed to make definitive remarks about causality

**Clustering Employee Absenteeism**

The goal of this algorithm (KMeans cluwas to potentially identify subtypes of datapoints. After these subtypes were identified, an attempt at explaining the found differences was made.

**Cluster 0 — "Structured & Seasonal Absentees"**Employees with **predictable, seasonal patterns** of absenteeism, typically seen in the colder months. Their absence reasons are slightly more non-medical or family-related.

**Key Characteristics:**

* Lower previous absences (16.3 vs 28.4)
* More absences in Q4 (October–December)
* High seasonal impact (Winter: 55%)
* Slightly longer commute distance (30.1 km)
* Slightly higher workload (261.7 vs 280.7)
* Higher social drinking (60%)
* More likely to have children (and more children when they do)

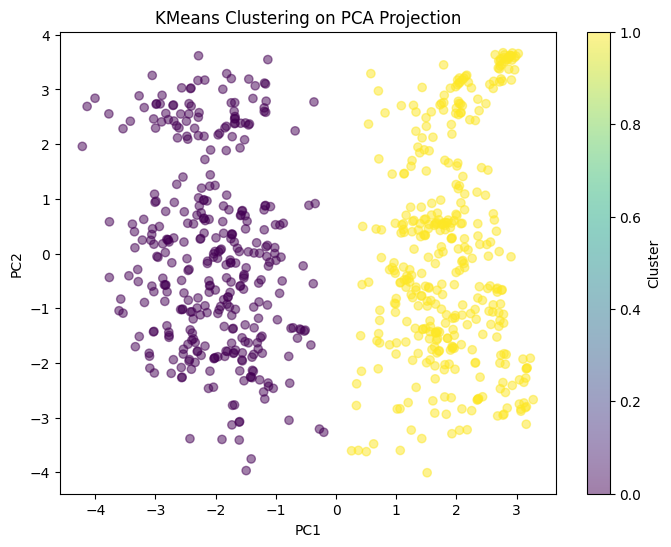
**Interpretation:**Structured employees whose absences may relate to predictable seasonal factors like flu or family responsibilities. Potential interventions include seasonal wellness initiatives or flexible holiday planning.

**Cluster 1 — "Routine & Frequent Absentees"**Employees with **frequent, short absences**, often tied to **routine medical or physical reasons**. They are more prevalent in recent data entries and show patterns early in the calendar year.

**Key Characteristics:**

* Higher previous absences (28.4 vs 16.3)
* More routine and physical health-related absences
* Higher representation in January–May
* Shorter commute distance (29.1 km)
* Slightly more social smoking
* More likely to come from 2008 data

**Interpretation:**This group may include employees managing chronic health conditions or those recently onboarded. Targeted interventions might include bundled medical appointments, telehealth options, or onboarding health programs.

 *Figure 4: Clustering split visualized*

**Disclaimer: Limitations of the Dataset and Analysis**

While this analysis provides insights into absenteeism patterns using the UCI Absenteeism at Work dataset, several limitations should be acknowledged:

* **Anonymized Data**  
  The dataset is partially anonymized and does not include identifiable demographic or occupational variables beyond basic attributes. This limits the contextual richness needed to fully explain absentee behavior (e.g., job type, health history, workplace culture).
* **Lack of Outcome Labels**  
  The dataset does not include clear target variables for supervised learning (e.g., productivity impact, return-to-work outcomes), which constrains the use of predictive models and limits the scope to exploratory or unsupervised techniques.
* **Data Quality & Balance**  
  Some reason categories (e.g., pregnancy-related, general medical) are underrepresented, and the "unknown" or zero-value reasons introduce noise. This may affect statistical robustness, especially for minor groups.
* **Temporal Relevance**  
  The dataset spans only three years (2007–2010) and may not reflect post-pandemic absenteeism trends, changes in remote work practices, or modern occupational health policies.
* **Simplified Feature Set**  
  Important factors like psychosocial stressors, organizational climate, and manager-employee dynamics are not captured, which are known to influence absenteeism.
* **Data size**  
  The dataset is of limited size (35 employees, 740 datapoints). This naturally limits the efficacy and generalizability of any model trained on this dataset. Furthermore, making subdivisions is practically impossible as it would divide the data in barely analysable chunks.